Mars Hyperspectral Data Processing using ICA and Bayesian Positive Source Separation

Hafrún Hauksdóttir[*], Saïd Moussaoui[†], Frédéric Schmidt[**], Christian Jutten[‡], Jocelyn Chanussot[‡] and David Brie[†]

[*]*University of Iceland, Reykjavik, Iceland*
[†]*Centre de Recherche en Automatique de Nancy (CRAN UMR CNRS 7039), France*
[**]*Laboratoire de Planétologie de Grenoble, France*
[‡]*Laboratoire des Images et des Signaux, Grenoble, France*

Abstract. The surface of Mars is currently being mapped with an unprecedented spatial resolution. This high resolution, and its spectral range, give the ability to pinpoint chemical species on Mars more accurately than before. The subject of this paper is to present a method to extract informations on chemicals using hyperspectral images. We propose to combine spatial Independent Component Analysis (ICA) [1] and spectral Bayesian Positive Source Separation (BPSS) [2]. The basic idea is to use spatial ICA yielding a rough classification of pixels, which allows selection of small, but relevant, number of pixels. BPSS is then applied for the estimation of the source spectra using this reduced set of pixels. Finally, the abundances of the components is assessed on the whole pixels of the images. Results of this approach are shown and evaluated by comparison with reference spectra.

INTRODUCTION

As solar light incident to a planetary surface is partially reflected back by interaction with the different constituents, the analysis of reflectance spectra may allow the identification and the quantification of the chemical species present at the surface of Mars. The simplest model assumes that each measured spectrum is a linear mixture of the component reflectance spectra, which is the case for geographical mixture of chemical species on the surface. The actual sources present significant cross-correlations and thus the fundamental assumption of independence in ICA is not satisfied. Therefore, ICA is not an adequate method for the unmixing [3]. Moreover, it is important to take into account the positivity constraint of both sources and mixing coefficients. This draws attention to Bayesian approach which is able to manage priors on both the sources and the mixing coefficients [4], such as positivity [2]. But, there we face the problem of high computation time when dealing with vast amount of data (more than 30.000 pixels). The first part of the paper is a short presentation of the data and the methods we used, which are Independent Component Analysis (ICA) and Bayesian Positive Source Separation (BPSS). The combination of the two approaches is described in more details and the results of applying the method on real data are discussed.

HYPERSEPCTRAL DATA MODELLING

The hyperspectral data are obtained with the OMEGA instrument (Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité), which is a spectrometer boarded on the European Space Agency Mars Express mission. The data set consists of 256 images in the infrared spectral region from 0.926 to 5.108 μm with a resolution of 0.014 μm roughly. The instrument has three channels, a visible channel and two infrared channels. The visible channel collects reemitted sunlight from the surface and atmosphere of Mars and the infrared channel collects thermal radiations given off from the surface. The spacecraft travels on an elliptical orbit which results in a resolution range from 300m to 4km. In each analyzed area, different minerals absorb and radiate different wavelengths of light. This basic physic allows us to identify the substances from the spectral information. Each sensor receives reflected light from an area that can be composed of several types of chemicals. The sunlight can be reflected from each material separately, which gives an additive mixture of spectra or it can be reflected from one material to another before reaching the sensor. The first case corresponds to the so called geometrical mixture while the second appears with intimate mixtures. In the sequel, we consider the case of geometrical mixtures.

Observation model

The observation model for a geometrical mixture can be expressed as [5]

$$L(x,y,\lambda) = \Phi(\lambda)\left(L_a(\lambda) + \sum_{p=1}^{P} \alpha_p(x,y) L_p(\lambda)\right) \cos\left[\theta(x,y)\right] \tag{1}$$

where $\Phi(\lambda)$ is the spectral atmospheric attenuation, $\theta(x,y)$ the angle between the solar vector and the surface normal, P the number of endmembers in the region of coordinates (x,y), $L_p(\lambda)$ the spectrum of the p-th endmember, $\alpha_p(x,y)$ its weight in the mixture and $L_a(\lambda)$ the radiation that did not arrive directly from the area under view. This mixture model can also be written as

$$L(x,y,\lambda) = \sum_{p=1}^{P} \alpha'_p(x,y) \cdot L'_p(\lambda) + E(x,y,\lambda) \tag{2}$$

where

$$\begin{cases} \alpha'_p(x,y) = \alpha_p(x,y)\cos\left[\theta(x,y)\right], \\ L'_p(\lambda) = \Phi(\lambda) L_p(\lambda), \\ E(x,y,\lambda) = \Phi(\lambda) L_a(\lambda)\cos\left[\theta(x,y)\right]. \end{cases}$$

As it can be seen in equation (2) the true endmember spectra are affected by the atmospheric absorption and the abundance fraction are corrupted by the solar angle effect. It is not attempt here to remove the atmospheric effect, and thus the spectra obtained are ideally the spectra of the endmembers with atmospheric absorption, but it

is attempted to correct the solar angle effect to determine the value of $\alpha_p(x,y)$, allowing a generation of a map of abundance $c_p(x,y)$ of the endmembers in the area. In addition, the plane of the solar angle effect is used for correction of the abundance fractions. An approximation of the solar angle effect can be calculated since the position of the sun and sensor is known at the time of the observation.

Mixing models

Let us consider a hyperpectral data cube with N_f images of $N_x \times N_y$ pixels obtained from N_f frequency bands. For simplicity, assume raw vectorized images $I_n(\lambda_k)$, with $1 \leqslant n = i + jN_y \leqslant N_x \times N_y$ (where i and j are the initial row and column image indices) is the *spatial index* and k, $k = 1,\ldots,N_f$, is the *spectral index* for the wavelength λ_k. Consequently, two representations of the hyperspectral data can be considered:

1. **Spectral mixture model:** each pixel of spatial index n gives an observed spectrum of N_f frequency samples, which is represented by the linear approximation

$$I_n(\lambda_k) \approx \sum_{p=1}^{N_c} a_{(p,n)} \psi_p(\lambda_k), \quad \forall n = 1,...,N_x \times N_y \tag{3}$$

   where $\psi_p(\lambda_k)$, for $p = 1,\ldots,N_c$, are the surface constituant reflectance spectra, and the number N_c is chosen according to the desired accuracy of the approximation. Denoting the vectorized image (of dimension $N_x \times N_y$) $\mathbf{I}(\lambda_k)$, the $(N_x \times N_y) \times N_c$ mixing matrix $\mathbf{A}$ and $\Psi(\lambda_k) = [\psi_1(\lambda_k),\ldots,\psi_{N_c}(\lambda_k)]^T$, this spectral mixture model is then expressed as:

$$\mathbf{I}(\lambda_k) \approx \mathbf{A} \cdot \Psi(\lambda_k). \tag{4}$$

2. **Spatial mixture model:** for each wavelength λ_k, the measured image $I_{\lambda_k}(n)$ is a weighted sum of N_c basis images, denoted $II_p(n)$, $p = 1,\ldots,N_c$. Denoting the set of observations (of size $N_f \times (N_x \times N_y)$) $\mathbf{I}_{\lambda_k}(n)$, each component of the model is:

$$I_{\lambda_k}(n) \approx \sum_{p=1}^{N_c} b_{(k,p)} II_p(n), \quad \forall k = 1,...,N_f, \tag{5}$$

   where $II_p(n)$ are independent images. In vector notations, denoting the $N_f \times N_c$ matrix $\mathbf{B}$ and $\mathbf{II}(n) = [II_1(n),\ldots,II_{N_c}(n)]^T$, one can write:

$$\mathbf{I}(n) \approx \mathbf{B} \cdot \mathbf{II}(n). \tag{6}$$

BLIND SOURCE SEPARATION

Source separation consists in retrieving unknown signals, $\mathbf{s} = (s_1(t),\ldots,s_n(t))^T$, which are observed through unknown mixtures of them. Denoting the observations $\mathbf{x}(t) = (x_1(t),\ldots,x_p(t))^T$, one can write:

$$\mathbf{x}(t) = \mathscr{F}(\mathbf{s}(t)), \tag{7}$$

where $\mathscr{F}(.)$ denotes the unknown mixture, a function from $\mathbb{R}^n$ to $\mathbb{R}^p$. If the number of observations p is greater then or equal to the number of sources, n, the main idea for separating the sources is to estimate a transform $\mathscr{G}(.)$ which inverses the mixture $\mathscr{F}(.)$, and provides estimates of the unknown sources. Of course, without other assumptions, this problem cannot be solved. Basically, it is necessary to have priors about

- the nature of the mixtures: it is very important to chose a separating transform $\mathscr{G}(.)$ suited to the mixture transform $\mathscr{F}(.)$,
- the sources: sources properties - even weak - are necessary for the estimation of $\mathscr{G}(.)$.

Because of the very weak assumptions, the problem is referred as blind source separation (BSS), and the method based on source independence property has been called independent component analysis (ICA) [1]. In the simplest case, the model is assumed to be linear and memoryless, *i.e.* $\mathscr{F}$ reduces to a mixing matrix $\mathbf{A}$ with scalar entries. This problem has been intensively studied in the last two decades, and many methods and algorithms are available, based on 4th-order statistics, entropic criteria, characteristic functions, etc. In the framework of hyperspectral data, although there is some evidence for a mixture model, one considers a sparse representation of the data using a sparse basis, with special properties. ICA provides such a model where the special property is mutual independence. The spatial mixture model (5) is considered. However, a scale indeterminacy can not be avoided when using ICA and without imposing positivity constraints on the independent components (ICs) or on the mixing matrix results in a sign ambiguity. An implementation of ICA used here is the well known JADE algorithm based on the joint approximate diagonalization of cumulant matrices [6].

On the other hand, in a Bayesian approach for source separation one can ideally incorporate every prior knowledge as long as this prior can be stated in statistical terms. As compared to ICA, in a Bayesian approach informations on the mixing coefficients are accounted during the analysis. The approach is founded on the likelihood $p(\mathbf{x}|\mathbf{A},\mathbf{s})$ and prior distributions of the source and mixing coefficients. Bayes' theorem leads to

$$p(\mathbf{A},\mathbf{s}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{A},\mathbf{s}) \times p(\mathbf{A}) \times p(\mathbf{s}) \tag{8}$$

From this posterior law, both the mixing matrix $\mathbf{A}$ and sources $\mathbf{s}$ can be estimated using various Bayesian estimators. A complete discussion on Bayesian approach to source separation is given in [4]. In the framework of spectral mixture data [2], the non-negativity of the sources and the mixing coefficients can be ensured using Gamma distributions as priors and the estimation of the sources and the mixing coefficients from the resulting posterior distribution is performed using Marginal posterior mean estimator and Markov Chain Monte Carlo (MCMC) methods.

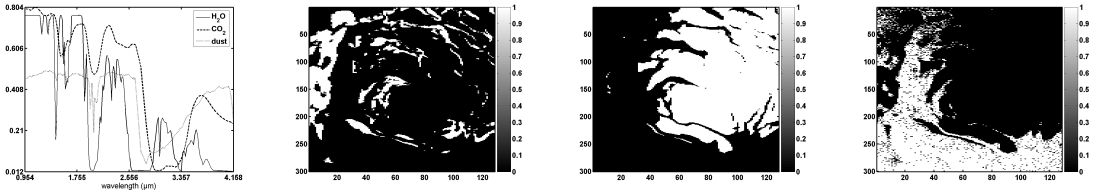
ICA AND BPSS APPLIED TO HYPERSPECTRAL DATA

The data used here covers the south polar region of Mars and are collected in the infrared region ranging from $0.96\mu m$ to $4.16\mu m$, total of 184 bands where noisy bands have been excluded. Some channels of the Omega instrument are badly corrupted by noise. These

channels are generally known or can be found with filtering. Throughout the paper, we illustrate our proposal using a benchmark data set which was also analyzed using other methods such as wavanglet classification

Reference data and ground truth

Two validate the results, two kinds of reference data are used (figure 1). First are the reference spectra of CO_2, H_2O and dust, which will be noted as $\psi_{CO_2}(\lambda)$, $\psi_{H_2O}(\lambda)$ and $\psi_{\text{dust}}(\lambda)$ respectively. The second are the results of a wavelet classification, which will be noted as $I_{CO_2}(n)$, $I_{H_2O}(n)$ and $I_{\text{dust}}(n)$. This classification is neither unique nor complete, that is pixels can be in more than one class and not all the pixels are classified.



(a) Reference spectra (b) H_2O (c) CO_2 (d) dust

FIGURE 1. (a) Reference spectra and (b,c,d) results of a wavelet classifier.

There are two main reasons why ICA is not effective by itself in the case of these hyperspectral images. The first reason is that the fundamental assumption of mutual independence is not fully satisfied. Both spatial and spectral independence are tested by looking at the covariance matrices $\mathbf{R}_s$ and $\mathbf{R}_\lambda$ of the ground truth images and the reference reflectance spectra, respectively :

$$\mathbf{R}_s(\text{dust}, CO_2, H_2O) = \begin{bmatrix} 1 & -0.61 & -0.24 \\ -0.61 & 1 & -0.25 \\ -0.24 & -0.25 & 1 \end{bmatrix}, \tag{9}$$

$$\mathbf{R}_\lambda(\text{dust}, CO_2, H_2O) = \begin{bmatrix} 1 & 0.58 & 0.88 \\ 0.58 & 1 & 0.73 \\ 0.88 & 0.73 & 1 \end{bmatrix}. \tag{10}$$

The second reason, which is a consequence of the first one, is that the independent components are found to be negative, which does not agree with the nature of the data. An interesting difference is seen between the spatial and spectral dimensions. A significant dependence can be seen between dust and CO_2 in the spatial dimension and between CO_2 and H_2O is in the spectral dimension. This dependence is also clearly demonstrated in the results of the ICA, since these elements are in fact not independent and therefore can not be effectively separated by ICA in most cases. When using spatial ICA, components of CO_2 and H_2O are retrieved but dust does not appear as a separate component. However, from the standpoint of spectroscopist it is of great interest to find the spectral signatures of the components, this is where the Bayesian

positive source separation approach comes to rescue. The additivity of spectra as seen in equation (2) leads naturally to assume the spectra to be sources, and therefore spectral BPSS is applied. Important properties of Baysian method are that it accounts for the positivity constraint and it avoids the need of orthogonality constraint, which appears to be essential after viewing the second order covariance of the spectra in equation (10). The prospect of having a abundance fraction is a great motivation for applying BPSS since alternative methods, like wavanglet classification, only give a classification of substances known to be in the region. However, a shortcoming of this method is the high computational cost due to the huge number of observations and to the use of MCMC methods.

Proposed Approach

Rather than using only one source separation approach, we propose to combine the two methods : JADE and BPSS. Spatial independent component analysis (with JADE) is used for selecting (a few number of) pixels in independent areas of the hyperspectral image. The mixture spectra provided by the selected pixels are then processed using BPSS to estimate the pure component spectra, with a lower computational cost.

Source number determination. A key point in blind source separation is the selection of the number of components. A traditional way to determine the number of sources is to look at the eigenvalues of the data covariance matrix. This method is not suitable in this case, since the sources do unfortunately not give strong directions for the covariance matrix. The method of selecting the strongest eigenvalues can also be treacherous in this case since specific chemicals can be located on a small area. Since the material is only projected on a few pixels, it does not have a great significance for the covariance matrix and could therefore be overlooked when choosing the strongest eigenvalues. The method proposed here is to determine the number of components from the spatial reconstruction SNR. This permits the detection of an area presented with only a few pixels, since components are found until all the image is well reconstructed. The threshold for the reconstruction is chosen around $20dB$, components are thus found until $SNR(n) > 20dB, \forall n$, where

$$SNR(n) = 10\log_{10}\left(\frac{\sum_{k=1}^{N_f} I_{\lambda_k}(n)^2}{\sum_{k=1}^{N_f}\left(I_{\lambda_k}(n) - \sum_{p=1}^{N_c} a_{\lambda_k,p} II_p(n)\right)^2}\right). \tag{11}$$

The spatial reconstruction information is both used for identification of the number of components for ICA and Bayesian approaches.

Selection of relevant pixels. The basic idea of the method used here is to use spatial ICA yielding a rough classification of pixels, which allows selection of small, but relevant, number of pixels. We define the relevant pixels associated to each source as those where the contribution of this source is important. To measure this contribution,

we use the spatial SNR lost which is also defined as the variation of the spatial SNR when one particular source is removed from the mixture. Thus

$$SNR_j(n) = SNR(n) - 10\log_{10}\left(\frac{\sum_{k=1}^{N_f} I_{\lambda_k}(n)^2}{\sum_{k=1}^{N_f}\left(I_{\lambda_k}(n) - \sum_{p=1,p\neq j}^{N_c} a_{\lambda_k,p} II_p(n)\right)^2}\right). \quad (12)$$

These SNRs are thresholded and the pixels selected at random locations from the resulting masks. The threshold used here is a simple Gaussian threshold with around 10% of the most relevant pixels. Each SNR lost image $SNR_j(n)$ is assumed to have a Gaussian distribution with mean μ_j and standard deviation σ_j and the threshold is chosen as:

$$T_j = \mu_j + 1.1 \times \sigma_j. \quad (13)$$

Abundance fraction determination. The goal of calculating abundance fractions is to find the amount of the endmembers in the region of each pixel. As seen in equation (1), the true endmember spectra are corrupted by the solar angle effect which is a function of the position (x,y). The measured spectra $I_n(\lambda) = L_\lambda$ of the mixture model (equation (3)) can be written as:

$$I_n(\lambda) = \sum_{p=1}^{N_c} (cos\left[\theta(x,y)\right]\alpha_p) \cdot L'_p(\lambda) \quad (14)$$

where atmospheric effect has been neglected, or can be viewed as inherent in the spectra $L_p(\lambda)$. The abundance fraction is the portion of each constituent in the geometrical mixture. From the mixture model in equation (3) and equation (14) it can be seen that the abundance fractions are not affected by the geometrical effet since:

$$c_p(x,y) = \frac{\alpha'_p}{\sum_{j=1}^{N_c}\alpha'_j} = \frac{\alpha_p \cos[\theta(x,y)]}{\sum_{j=1}^{N_c}\alpha_j \cos[\theta(x,y)]} = \frac{\alpha_p}{\sum_{j=1}^{N_c}\alpha_j}. \quad (15)$$

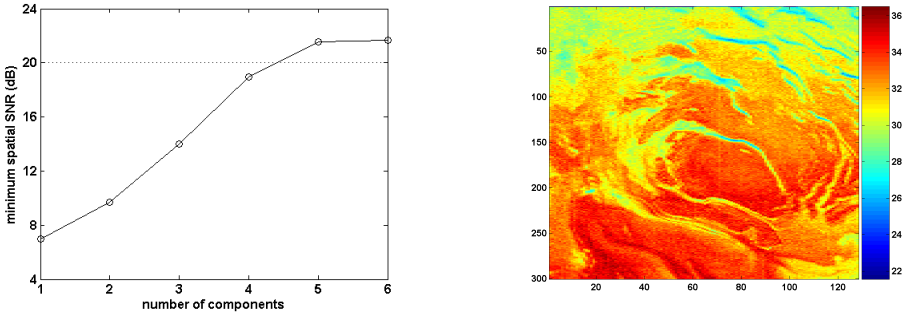
APPLICATION RESULTS

We present the results of the application of the proposed approach to the Omega hyperspectral images obtained from the south polar cap of Mars by the Omega/Mars Express.

Spatial independent component analysis. Figure 2 illustrates the variation of the reconstruction error with respect to the number of sources. The condition that all pixels should be reconstructed with a minimum of $20dB$ for the spatial SNR was reached when 5 components were used. The 5 Independent components found by using JADE can be seen on figures 3 and 4. It can be noted that the estimated mixing coefficients present both positive and negative values, which means that these quantities can not be compared with the pure spectra of chemical species.

Using these ICA results, one can calculate the spatial SNR lost for each independent components and select randomly 50 pixels from each class using the Gaussian threshold.

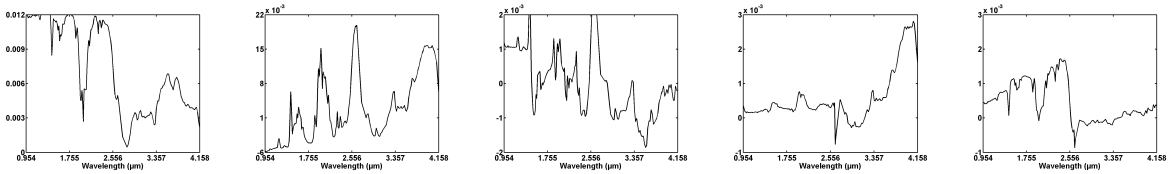
Figure 5 shows these spatial SNRs and the randomly selected pixels. It can be seen that pixels are chosen from the expected locations of *dust*, H_2O and CO_2 ices.



(a) Minimum spatial SNR vs. number of components

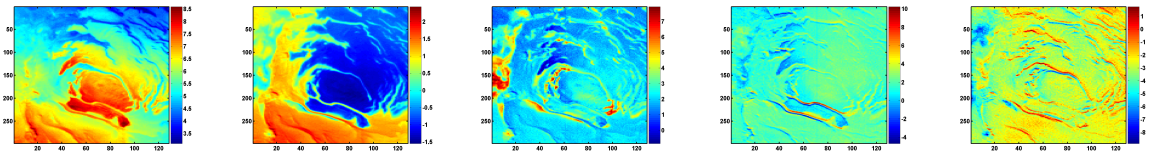
(b) Spatial SNR for 5 sources

FIGURE 2. Spatial Reconstruction error for different number of sources.



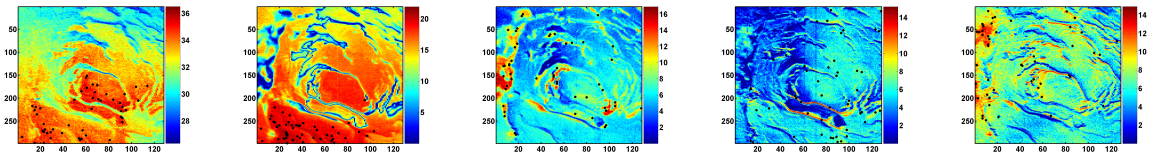
(a) Component 1 (b) Component 2 (c) Component 3 (d) Component 4 (e) Component 5

FIGURE 3. Estimated mixing coefficient (MCs) profiles using JADE



(a) Component 1 (b) Component 2 (c) Component 3 (d) Component 4 (e) Component 5

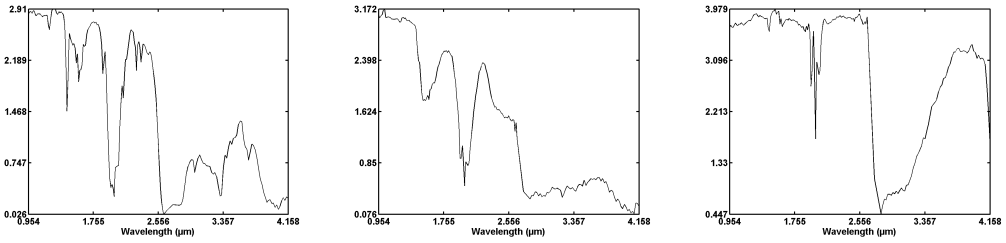
FIGURE 4. Estimated independent components (ICs) using JADE



(a) Component 1 (b) Component 2 (c) Component 3 (d) Component 4 (e) Component 5

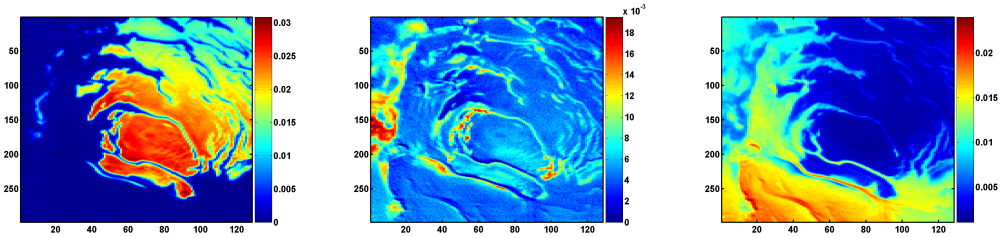
FIGURE 5. Spatial SNR lost for different components. The circles indicate the randomly selected pixels

Spectral Bayesian separation with positivity constraints. The results of the Bayesian separation from the selected pixels are shown is figures 6 and 7. To handle the scale ambiguity in source separation, the estimated pure spectra are normalized to have unit variance.



(a) Source 1 (b) Source 2 (c) Source 3

FIGURE 6. The pure spectra found by BPSS



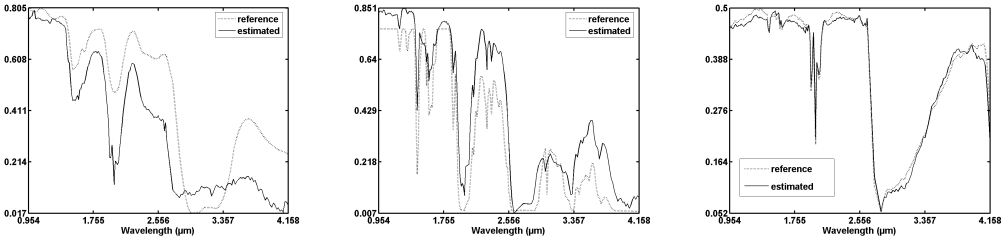
(a) Endmember 1 (b) Endmember 2 (c) Endmember 3

FIGURE 7. Mixing coefficients estimated using BPSS

The Identification of the spectra is straight forward from the correlation with the reference spectra as seen from equation (16) where $\mathbf{R} = [r\{\psi_i, \hat{\psi}_j\}]$ is the correlation between the reference spectra $\Psi_j = \{\psi_{H_2O}, \psi_{CO_2}, \psi_{dust}, \}$ and the estimated pure spectra $\hat{\Psi}_i = \{\hat{\psi}_1, \hat{\psi}_2, \hat{\psi}_3\}$.

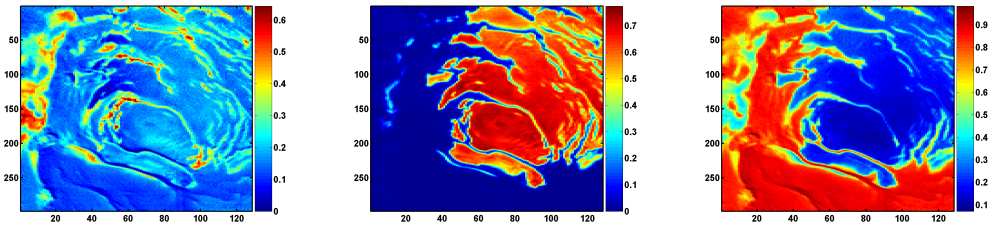
$$\mathbf{R} = \begin{bmatrix} 0.79 & \mathbf{0.91} & 0.87 \\ \mathbf{0.96} & 0.89 & 0.55 \\ 0.65 & 0.72 & \mathbf{0.99} \end{bmatrix} \tag{16}$$

After scaling and permutation of the identified pure spectra, the reference spectra are plotted together on figure 8. It can be noted the similarity between the estimated spectra and the true ones. On the other hand, one can also compute the abundance fractions in each pixel of the hyperspectral images, which are shown in figures 9. One can also see that the abundance fraction not only agree with the results of the wavangelet classification shown in figure 1, but also give a more precise information on the amount of the chemicals in the observed region of Mars.



(a) H_2O (b) CO_2 (c) dust

FIGURE 8. Comparison of reference and estimated pure spectra



(a) H_2O (b) CO_2 (c) dust

FIGURE 9. Estimated abundance fractions

CONCLUSION

In this paper, we have presented an application of independent component analysis and Bayesian source separation to the processing of hyperspectral data. This approach allowed the extraction of relevant information regarding the observed region from Mars. We particularly, determined the pure spectra of H_2O, CO_2 and dust ices as well as the abundance of these chemicals. Future works concern the processing of other regions of this planetary where other expected and unexpected chemicals are present.

REFERENCES

1. Comon, P., *Signal Processing*, **36**, 287–314 (1994).
2. Moussaoui, S., Brie, D., Carteret, C., and Djafari, A. M., "Application of Bayesian Non-negative Source Separation to Mixture Analysis in Spectroscopy," in *24th International Workshop MaxEnt*, edited by R. Fischer, R. Preuss, and U. von Toussaint, AIP Conference Proceedings, Garching, Germany, 2004, vol. 735, pp. 237–244.
3. Hauksdottir, H., Jutten, C., Schmidt, F., Chanussot, J., Benediktsson, J., and Douté, S., "The Physical Meaning of Independent Components and Artifact Removal of Hyperspectral Data from Mars using ICA," in *7th Nordic Signal Processing Symposium (NORSIG'2006)*, Reykjavik, Iceland, 2006.
4. Mohammad-Djafari, A., "A Bayesian approach to source separation," in *19th International Workshop MaxEnt*, edited by J. T. Rychert, G. J. Erickson, and C. R. Smith, AIP Conference Proceedings, Boise, Idao, 1999, vol. 567, pp. 221–244.
5. Healey, G., and Slater, D., *IEEE Transactions on Geoscience and Remote Sensing*, **37** (1999).
6. Cardoso, J.-F., and Souloumiac, A., *IEE Proceedings-F*, **140**, 362–370 (1993).